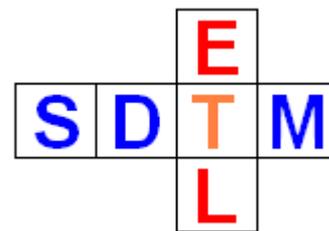


SDTM-ETL 5.0



Tutorial: Using ChatGPT and Word Similarity for mapping suggestions

Author: Jozef Aerts, XML4Pharma

Last update: 2025-03-11

Table of Contents:

Introduction.....	1
Using ChatGPT for mapping suggestions.....	1
Using Word Similarity and CDISC Controlled Terminology	5
Conclusions.....	10

Introduction

Independent of what tool is used, mapping collected clinical or pre-clinical data to SDTM or SEND requires a lot of experience, especially when the study design and/or CRFs does not use the [CDISC Controlled Terminology \(CDISC-CT\)](#) and/or the CRFs are not (electronically, as can be done on ODM) annotated with the SDTM information.

One source for being able to answer questions like "to what SDTM/SEND domain and variable do I map the datapoint XYZ too", with "XYZ" e.g. being "WBC", or "oxygen saturation", surely are the by CDISC published [Implementation Guides](#). Another very good source is to go through the CDISC-CT files themselves, which can however be cumbersome. And nowadays, there is Artificial Intelligence (AI) like ... [ChatGPT](#) ...

Using ChatGPT for mapping suggestions

ChatGPT (based on artificial intelligence - AI) has, for many of us, become part of our daily life. Especially for SDTM beginners, it can provide reasonable hints for mappings. Therefore, we have build an interface to ChatGPT into SDTM-ETL starting from v.4.2.

Its use however requires that the user has obtain a [ChatGPT API key](#), which needs to be added to the "properties.dat" file (see further on).

In order to use ChatGPT for obtaining a mapping hint, first select an item in the ODM tree, for example "WBC":

? The following question regarding your ODM/SDTM item will be asked to ChatGPT

To what CDISC SDTM domain should I map Leukocytes to?

SDTM Genius ChatGPT Regular ChatGPT

Ask ChatGPT!

ChatGPT's answer:

OK

As of SDTM-ETL v.5.0, you can choose between an SDTM-specialized extension to ChatGPT named "Genius SDTM" and the regular ChatGPT. We recommend to use the first one, as it provides considerably better suggestions. It can also be accessed in any modern browser at:

<https://chatgpt.com/g/g-67cf79284c8c8191a3f47b870ef8878f-sdtm-genius>

In the above textfield, one can of course than still change the wording of the question. Clicking "Ask ChatGPT" then leads (after a few seconds) to:

? The following question regarding your ODM item will be asked to ChatGPT

To what CDISC SDTM domain should I map WBC to?

Ask ChatGPT!

ChatGPT's answer:

The CDISC SDTM domain to which WBC (white blood cell count) should be mapped is the Laboratory

OK

with ChatGPT's answer: *"The CDISC SDTM domain to which WBC (white blood cell count) should be mapped is the Laboratory (LB) domain."*

However, when one changes the question into: "To what CDISC SDTM domain and SDTM variable should I map WBC to?", ChatGPT's answer may not be entirely correct.

For example, some time ago, we got:

J You

To what CDISC SDTM domain and SDTM variable should I map WBC to

ChatGPT

In CDISC (Clinical Data Interchange Standards Consortium) SDTM (Study Data Tabulation Model), laboratory data such as White Blood Cell count (WBC) is typically mapped to the "LB" (Laboratory) domain. Specifically, WBC would be mapped to the variable named "LBCF" (Laboratory Test Result Finding).

Here's an example of how the mapping might look:

- SDTM Domain: LB (Laboratory)
- SDTM Variable: LBCF (Laboratory Test Result Finding)

as there is no SDTM variable "LBCF". The answer should be "LBORRES", and some information could be provided about the use of LBSTRESN and LBSTRESC.

We do see however that ChatGPT is becoming better every day, also for clinical research and for mapping to SDTM and SEND, as there is already so much knowledge available in articles, forum discussions and blogs. "SDTM Genius" does even better, and is still evolving.

To be able to work with ChatGPT from within SDTM-ETL, one will need to have a ChatGPT account and API-key and copy the latter into the "properties.dat" file. For example:

properties.dat - Editor

Datei Bearbeiten Format Ansicht Hilfe

```
language=en
languagefixed=true
# logfilepath=C:\temp
loglevel=DEBUG
sasviewerlocation=C:\Program Files\SAS Institute\SAS System Viewer\Sv.exe
adobereaderlocation="C:\Program Files\Adobe\Acrobat DC\Acrobat\Acrobat.exe"
# CDISC Library API key
cdisclibraryapikey=f[REDACTED]
# ChatGPT API key (without "Bearer")
chatgptapikey=sk-[REDACTED]
# other settings
```

As of SDTM-ETL v.4.4, you can also add two other ChatGPT parameters in the "properties.dat" file. These are:

- "chatgptmodel": the model to be used by ChatGPT (default: "gpt-3.5-turbo")
- "chatgptmaxtokens": the maximal number of tokens (see the [API documentation](#))

REMARK: Depending on your ChatGPT plan, the number of requests per hour may be limited. For more information on these limitations, see <https://platform.openai.com/docs/guides/rate-limits/usage-tiers>

Using Word Similarity and CDISC Controlled Terminology

CDISC-CT codelists contain almost¹ all allowed values for SDTM/SEND variables that are under controlled terminology. In many cases, the codelists also contain at least some information about what domain and variable the codelists is meant to be used for.

For example from the [CDISC Library Browser](#):

Codelist	Extensible	Name	Submission Value	Synonyms	Definition	NCI Preferred ...
C160922	Yes	Laboratory Analytical Method Calculation Formula	LBANMET	Laboratory Analytical Method Calculation Formula	Terminology relevant to the formula names associated with a laboratory test calculation.	CDISC SDTM Laboratory Analytical Method Calculation Formula Terminology
C65047	Yes	Laboratory Test Code	LBTESTCD	Laboratory Test Code	Terminology used for laboratory test codes of the CDISC Study Data Tabulation Model.	CDISC SDTM Laboratory Test Code Terminology
C67154	Yes	Laboratory Test Name	LBTEST	Laboratory Test Name	Terminology used for laboratory test names of the CDISC Study Data Tabulation Model.	CDISC SDTM Laboratory Test Name Terminology
C102580	Yes	Laboratory Test Standard Character Result	LBSTRESC	Laboratory Test Standard Character Result	Terminology relevant to the standardized results, in character format, for laboratory tests.	CDISC SDTM Laboratory Test Standard Character Result Terminology

In some other cases, this information is (unfortunately) not provided:

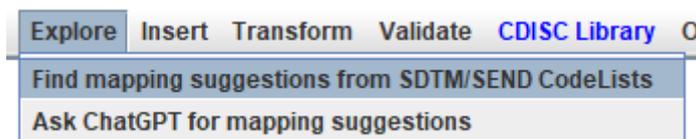
SDTM Controlled Terminology Package 55 Effective 2023-09-29

[Export](#)

Codelist	Extensible	Name	Submission Value	Synonyms	Definition	NCI Preferred ...
C141657	No	10-Meter Walk/Run Functional Test Test Code	TENMW1TC	10-Meter Walk/Run Functional Test Test Code	10-Meter Walk/Run test code.	CDISC Functional Test 10-Meter Walk/Run Test Code Terminology
C141656	No	10-Meter Walk/Run Functional Test Test Name	TENMW1TN	10-Meter Walk/Run Functional Test Test Name	10-Meter Walk/Run test name.	CDISC Functional Test 10-Meter Walk/Run Test Name Terminology

From the definition and description here, one cannot see that the domain is QS and the variable QSTESTCD.

Using the menu "Explore - Find Mapping Suggestions from SDTM/SEND CodeLists"



one can search the codelists for certain entries. For example, after having selected "WBC" in the ODM tree, and using the menu, a dialog will be displayed:

¹ Sometimes, for variables that are under controlled terminology, the CT (unfortunately) does not contain all allowed values. Examples of such values are "OTHER", "MULTIPLE", "xxALL" (e.g. LBALL).

The system will now try to find a match between the provided ODM name and the --TESTCD/--TEST codelists for SDTM.

Matches will be sorted by word similarity and provide suggestions for a suitable domain to which the ODM item can be mapped.

ODM Name:

Whole word

Limit to:

Find mapping suggestions

Results of comparison - % similarity

Close

showing additional features. These will be explained later on.

When one then clicks the button "Find mapping suggestions", the system will start looking into the currently loaded CDISC Controlled Terminology", as well as in the definitions as in the synonyms, and after some time, show a ranking which based on "word similarity"² (case-insensitive):

² The algorithm uses the "Levenshtein distance" - see e.g. https://en.wikipedia.org/wiki/Levenshtein_distance

Suggested variables for ODM Name WBC

Domain	Variable	Value	Label	Similarity %
CP	CPTTESTCD	WBC	Leukocytes	100.0
LB	LBTESTCD	WBC	Leukocytes	100.0
CP	CPTESTCD	RBC	Erythrocytes	66.7
LB	LBTESTCD	RBC	Erythrocytes	66.7
CP	CPTESTCD	WBCCE	Leukocytes/Total Cells	60.0
LB	LBTESTCD	CSWBC	WBC Casts	60.0
LB	LBTESTCD	WBCCE	Leukocytes/Total Cells	60.0
LB	LBTESTCD	ALBC	Albumin Clearance	50.0
LB	LBTESTCD	DGNWBC	Degenerated Leukocytes	50.0
LB	LBTESTCD	HGBC	Hemoglobin C	50.0
LB	LBTESTCD	IBCT	Total Iron Binding Capacity	50.0
LB	LBTESTCD	IBCU	Unsaturated Iron Binding Cap...	50.0
LB	LBTESTCD	VBCE	Viable Cells	50.0
LB	LBTESTCD	WBCCLMP	Leukocyte Cell Clumps	42.9
LB	LBTESTCD	WBCDIFF	Leukocyte Cell Differential	42.9
LB	LBTESTCD	ABNCE	Abnormal Cells	40.0
LB	LBTESTCD	CSWBC	WBC Casts	40.0

showing that "WBC" is a valid value for LBTESTCD in the LB domain and for CPTTESTCD in CP (Cell Phenotype Findings) with the label (i.e. the value in --TEST) being "Leukocytes". Interesting is e.g. when one searches for "oxygen saturation":

Suggested variables for ODM Name oxygen saturation

Domain	Variable	Value	Label	Similarity %
LB	LBTESTCD	OXYSAT	Oxygen Saturation	100.0
VS	VSTESTCD	OXYSAT	Oxygen Saturation	100.0
	RETESTCD	VO2	Oxygen Consumption	66.7
LB	LBTESTCD	TFRRNSAT	Transferrin Saturation	59.1
LB	LBTESTCD	OXYCAP	Oxygen Capacity	58.8
	ADCTC	ADCOR	Orientation	52.9

i.e. "oxygen saturation" can be mapped to either "LB" or "VS". Very curious!

The reason behind this is that when the oxygen saturation is measured in the lab, using a blood sample, it goes into LB. When however, it is measured using a pulse-oximeter, it goes into VS. Both measure the same property. Logical, isn't it?

SDTM is (still) about categorization (post-coordination). For "Findings", essentially --LOINC should be the identifier, not --TESTCD.

The search for a suitable domain and variable may take several minutes (time for a cup of coffee or tea ...), but one can speed up the process by only looking for "exact" words (i.e. the word provided must be a substring of the word in the CT), and using the field "Limit to", allowing to limit the search to specific codelists. For example for "oxygen saturation", only looking into codelists that are about "laboratory", the result will be:

Suggested variables for ODM Name oxygen saturation



Domain	Variable	Value	Label	Similarity %
LB	LBTESTCD	OXYSAT	Oxygen Saturation	100.0
LB	LBTESTCD	SAO2FIO2	Oxygen Saturation/Fraction Inspired O2	44.7

without any mentioning VS at all, but running much faster.

Now, let's search for "stair", as we have e.g. a functional test about number of stairs that the patient can run in a specific time. We want to limit to codelists that have "functional test" in their name or description, so we use:

The system will now try to find a match between the provided ODM name and the --TESTCD/--TEST codelists for SDTM.

Matches will be sorted by word similarity and provide suggestions for a suitable domain to which the ODM item can be mapped.

ODM Name:

Whole word

Limit to:

Find mapping suggestions

Results of comparison - % similarity

Domain	Variable	Value	Label	Similarity %
	A4STR1TC	A4STR102	A4STR1-Time to Do 4-Stair Ascend	15.6
	D4STR1TC	D4STR102	D4STR1-Time to Do 4-Stair Descend	15.2
	A4STR1TC	A4STR101	A4STR1-Was 4-Stair Ascend Performed	14.3
	D4STR1TC	D4STR101	D4STR1-Was 4-Stair Descend Performed	13.9

and the obtained result is:

Suggested variables for ODM Name stair



Domain	Variable	Value	Label	Similarity %
	A4STR1TC	A4STR102	A4STR1-Time to Do 4-Stair Ascend	15.6
	D4STR1TC	D4STR102	D4STR1-Time to Do 4-Stair Descend	15.2
	A4STR1TC	A4STR101	A4STR1-Was 4-Stair Ascend Performed	14.3
	D4STR1TC	D4STR101	D4STR1-Was 4-Stair Descend Performed	13.9

as one sees, no domain is provided, as the codelist does not provide information about what domain the codelist can be used. Also the "variable" provided is not a real variable, but simply the codelist name.

In the CDISC Library Browser:

Packages

DDF CT 2023-09-29 Protocol CT 2023-09-29 **SDTM CT 2023-09-29** SEND CT 2023-09-29

← C141663

Extensible: No
Submission Value: A4STR1TC
Definition: 4-Stair Ascend test code.
NCI Preferred Term: CDISC Functional Test 4-Stair Ascend Test Code Terminology
Synonyms: 4-Stair Ascend Functional Test Test Code

Filter results

Term	Submission Value	Synonyms	Definition	NCI Preferred Term
C174103	A4STR101	A4STR1-Was 4-Stair Ascend Performed	4-Stair Ascend - Was the 4-stair ascend performed?	4-Stair Ascend - Was 4-Stair Ascend Performed
C141706	A4STR102	A4STR1-Time to Do 4-Stair Ascend	4-Stair Ascend - If yes, time taken to do 4-stair ascend.	4-Stair Ascend - Time to Do 4-Stair Ascend
C147590	A4STR103	A4STR1-Wear Orthoses	4-Stair Ascend - If yes, did subject wear orthoses?	4-Stair Ascend - Wear Orthoses
C141707	A4STR104	A4STR1-Test Grade	4-Stair Ascend - Test grade.	4-Stair Ascend - Test Grade

Although some information is obtained about the possible codes for the mapping, the CodeList itself does not provide the information to what domain and variables the "stair" tests need to be mapped too, which is a pity.

When we ask "regular" ChatGPT, we get:

to what CDISC SDTM domain and variable must I map functional tests about stair ascending?

In CDISC SDTM (Study Data Tabulation Model), functional tests related to stair ascending (such as a timed stair ascent test) typically belong to the FT (Functional Tests) domain.

For specific variable mapping:

- **Domain:** FT (Functional Test)
- **Test Name (FTTEST):** "stair Ascending Test" (or a specific name that reflects the test performed)
- **Test Code (FTTESTCD):** A short code, such as "STAIRASC"
- **Result (FTSTRESC / FTSTRESN):** The result of the test, which can be categorical (FTSTRESC) or numerical (FTSTRESN)
- **Result Unit (FTSTRESU):** If a time-based test, this could be "seconds" (e.g., time taken to ascend a flight of stairs)
- **Position (FTPOS):** If relevant, specify body position (e.g., "standing")
- **Category (FTCAT):** "Mobility" or another appropriate category
- **Visit (VISITNUM, VISIT):** Indicating when the test was performed
- **Timing Variables:** FTDTCT for date/time of assessment

Would you like help with controlled terminology or example datasets?

which is pretty good ...

Conclusions

Although still in its infancy for the case of clinical research, ChatGPT can be used for getting mapping suggestions for very simple cases, e.g. for lab tests, vital signs tests etc.. For more complicated cases, like functional tests, it might still provide wrong answers.

Especially with "specializations" like "SDTM Genius" it is however getting better every day and learning fast.

An interesting article from my colleague and friend Angelo Tinnazi about ChatGPT CDISC knowledge can be found [here](#).

Searching through the loaded CDISC codelists can provide good hints about to what domain and variable a specific source data point must be mapped. This however only works well when the codelist name and/or description contains information about the usage of the codelist, which is not always present.